Introduction

Information Systems

Big Data Analytics

Presented by: Dr Sherin El Gokhy

Basic Methods

Module 3 – Review of Basic Data Analytic Methods Using R

Module 3: Review of Basic Data Analytic Methods Using R

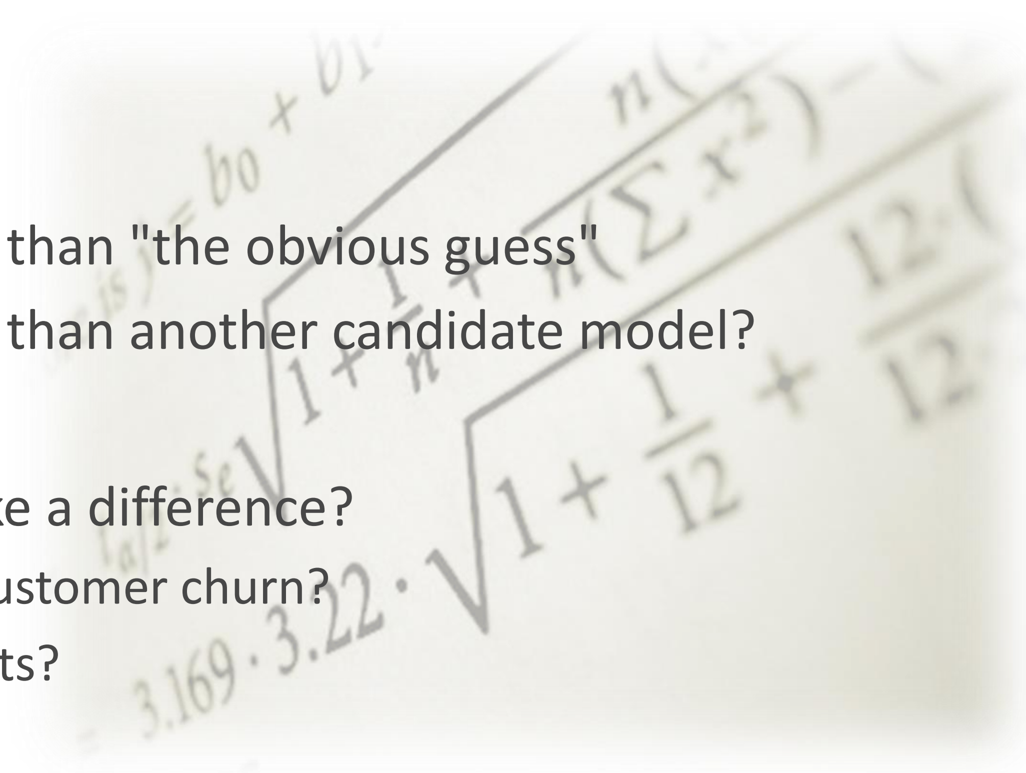
Part 3: Statistics for Model Building and Evaluation

During this lesson the following topics are covered:

- Statistics in the Analytic Lifecycle
- Hypothesis Testing
- Difference of means
- Significance, Power, Effect Size
- ANOVA
- Confidence Intervals

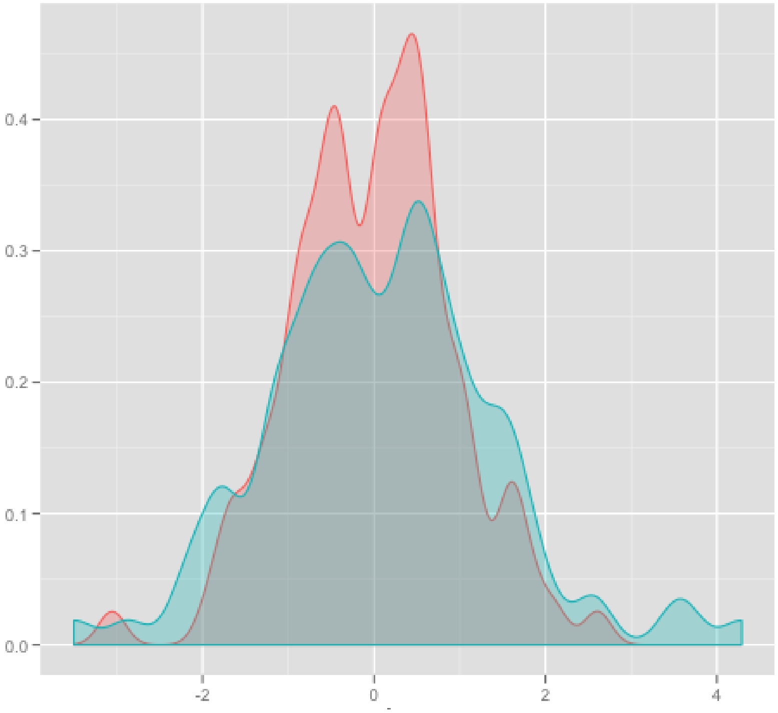
Statistics in the Analytic Lifecycle

- Model Building and Planning
  - Can I predict the outcome with the inputs that I have?
  - Which inputs?
- Model Evaluation
  - Is the model accurate?
  - Does it perform better than "the obvious guess"
  - Does it perform better than another candidate model?
- Model Deployment
  - Do my predictions make a difference?
    - Are we preventing customer churn?
    - Have we raised profits?



Hypothesis Testing

- Fundamental question: "Is there a difference between the populations based on samples?“
  - Examples : Mean, Variance



Variance: a measure of how data points differ from the mean

- Data Set 1: 3, 5, 7, 10, 10
- Data Set 2: 7, 7, 7, 7, 7

What is the mean and median of the above data set?

Data Set 1: mean = 7, median = 7

Data Set 2: mean = 7, median = 7

But we know that the two data sets are not identical! The **variance** shows how they are different.

We want to find a way to represent these two data set numerically.

How to Calculate?

- We estimate the spread of a distribution as the extent to which the values in the distribution differ from the mean and from each other.

$$\frac{\sum(x - \bar{X})}{N}$$

- The average of the squared deviations about the mean is called the <u>variance</u>.

$$s^2 = \frac{\sum (x - \overline{X})^2}{n}$$

- The standard deviation s is the **square root** of the **Variance.**

Example

	Score X	$X - \overline{X}$	$(X - \overline{X})^2$
1	3		
2	5		
3	7		
4	10		
5	10		
Totals	35		

The mean is 35/5=7.

Example(continued)

	Score X	$X - \overline{X}$	$(X - \overline{X})^2$
1	3	3-7=-4	
2	5	5-7=-2	
3	7	7-7=0	
4	10	10-7=3	
5	10	10-7=3	
Totals	35		

Example(continued)

	Score X	$X - \overline{X}$	$(X - \overline{X})^2$
1	3	3-7=-4	16
2	5	5-7=-2	4
3	7	7-7=0	0
4	10	10-7=3	9
5	10	10-7=3	9
Totals	35		38

Example(continued)

	Score X	$X-\bar{X}$	$(X-\bar{X})^2$
1	3	3-7=-4	16
2	5	5-7=-2	4
3	7	7-7=0	0
4	10	10-7=3	9
5	10	10-7=3	9
Totals	35		38

$$s^2 = \frac{\sum\left(x-\bar{X}\right)^2}{n} = \frac{38}{5} = 7.6$$

Example2

Dive	Mark	Myrna
1	28	27
2	22	27
3	21	28
4	26	6
5	18	27

Find the mean, median, range?

mean	**23**	**23**
median	**22**	**27**
range	**10**	**22**

Which diver was more consistent?

Dive	Mark's Score X	$X - \overline{X}$	$(X - \overline{X})^2$
1	28	5	25
2	22	-1	1
3	21	-2	4
4	26	3	9
5	18	-5	25
Totals	115	0	64

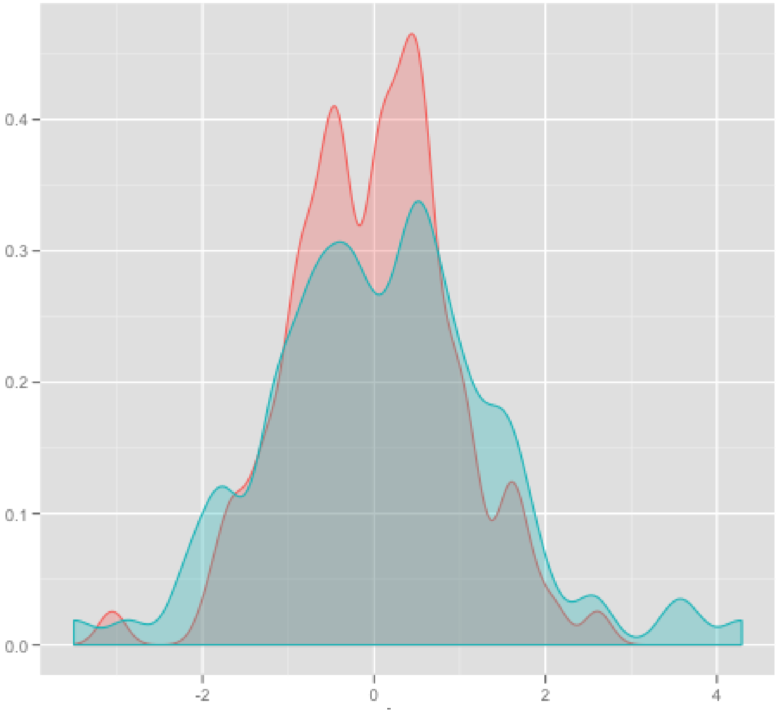
Mark’s Variance = 64 / 5 = 12.8
Myrna’s Variance = 362 / 5 = 72.4

Conclusion: Mark has a lower variance therefore he is more consistent.

Hypothesis Testing is a common technique to assess the difference or significance

- Null hypothesis : There is no difference

- Alternate hypothesis : There is a difference

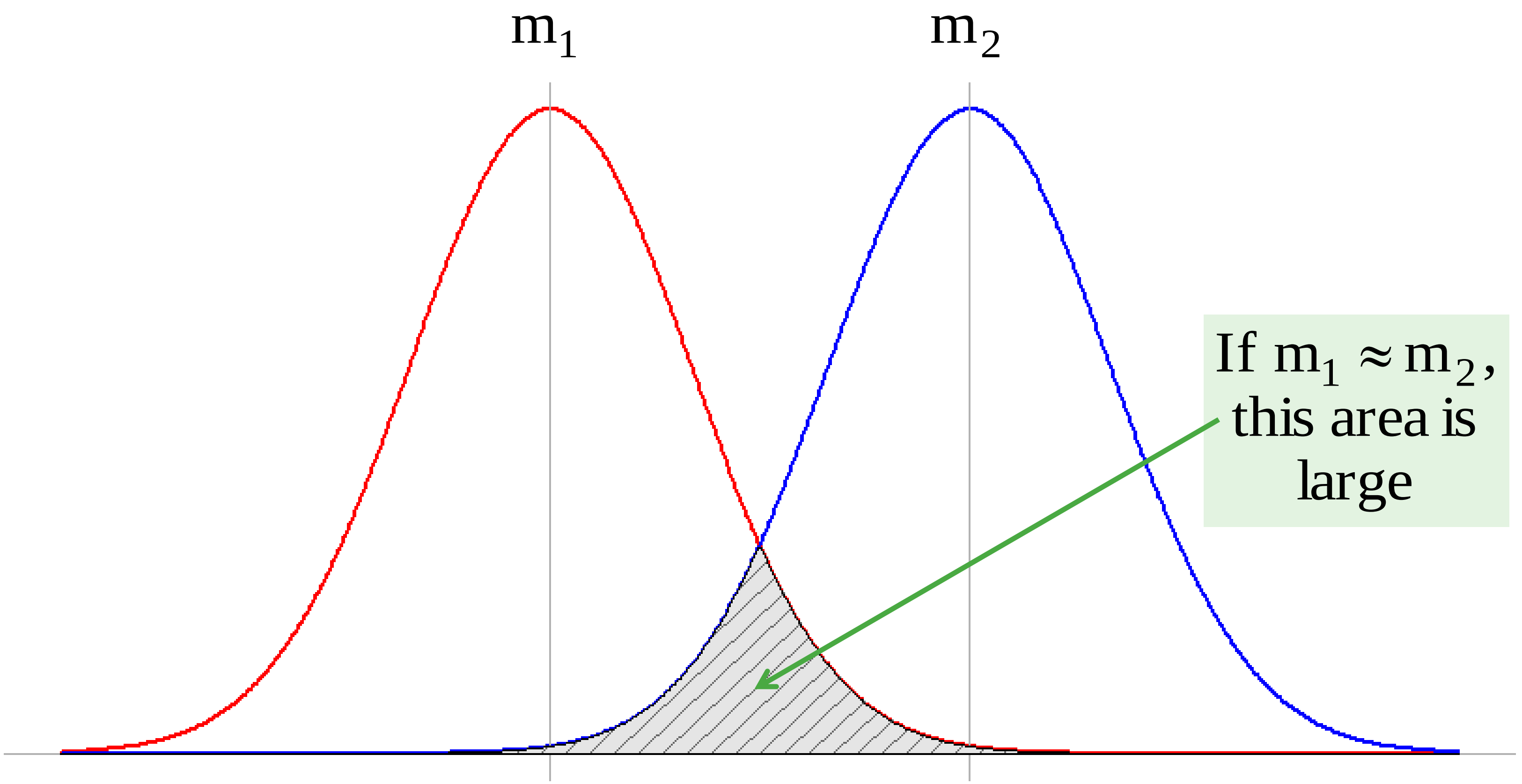


- The central aim of statistical test is to determine the likelihood of a value in a sample **under the assumption that the Null hypothesis is true**
  - The H0 states that there is no statistically significant difference between your sample and a reference population (or between two samples)
  - The H1 states the opposite, i.e. that there is a statistically significant difference between your sample and a reference population (or between two samples)

Null and Alternative Hypotheses: Examples

Null Hypothesis	Alternative Hypothesis
The average squared prediction error from the model is the same as the average squared prediction error from the null model.	The model predicts better than the null model: • The average squared prediction error from the model is smaller than that of the null model
This variable does not affect the outcome: • The coefficient value is zero	The variable does affect outcome: • Coefficient value is non-zero
The model predictions do not improve revenue(income): the same with or without intervention of hypothesis	Interventions based on model predictions improve revenue: • A/B Testing, ANOVA

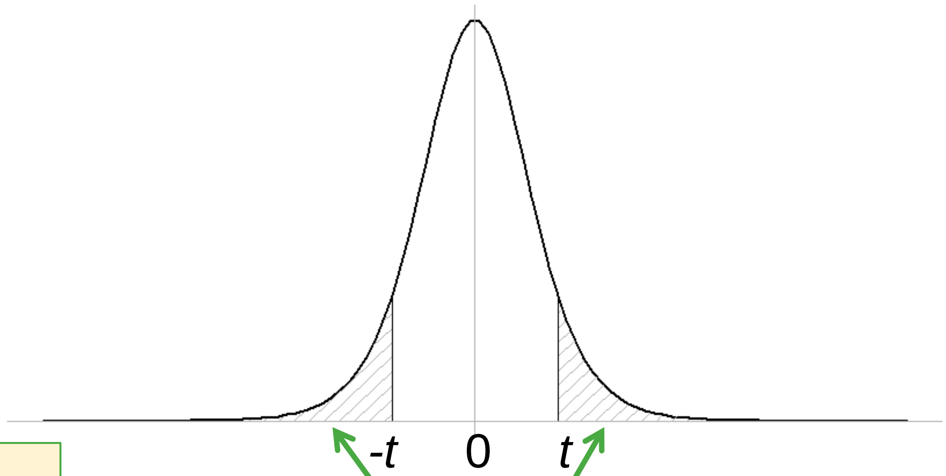
Intuition: Difference of Means



Welch's t-test

t-statistic: $$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

(this is the t-statistic for the Welch t-test)




```
> x = rnorm(10) # distribution centered at 0
> y = rnorm(10,2) # distribution centered at 2
> t.test(x,y)

        Welch Two Sample t-test

data:  x and y
t = -7.2643, df = 15.05, p-value = 2.713e-06
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -2.364243 -1.291811
sample estimates:
mean of x mean of y
0.5449713 2.3729984
```

p-value: area under the tails of the appropriate student's distribution

if p-value is small (say < 0.05), then reject the null hypothesis and assume that $m_1 <> m_2$

m_1 and m_2 are "significantly different"

Wilcoxon Rank Sum Test

- t-test assumes that the populations are normally distributed
  - Sometimes this is close to true, sometimes not
- Wilcoxon Rank Sum test
  - Makes no assumption about the distributions of the populations
  - More robust test for difference of means
  - if p-value is small: reject the null hypothesis (equal means)

```
> mean(x)
[1] 0.5449713
> mean(y)
[1] 2.372998
> wilcox.test(x, y)

            wilcoxon rank sum test

data:  x and y
W = 2, p-value = 4.33e-05
alternative hypothesis: true location shift is not equal to 0
```

Wilcoxon Rank Sum Test

Let N be the sample size, i.e., the number of pairs. Thus, there are a total of 2*N* data points. For pairs $i = 1, \ldots, N$, let $x_{1,i}$ and $x_{2,i}$ denote the measurements.

H_0: difference between the pairs follows a symmetric distribution around zero

H_1: difference between the pairs does not follow a symmetric distribution around zero.

1. For $i = 1, \ldots, N$, calculate $|x_{2,i} - x_{1,i}|$ and $\operatorname{sgn}(x_{2,i} - x_{1,i})$, where sgn is the sign function.
2. Exclude pairs with $|x_{2,i} - x_{1,i}| = 0$. Let N_r be the reduced sample size.
3. Order the remaining N_r pairs from smallest absolute difference to largest absolute difference, $|x_{2,i} - x_{1,i}|$.
4. Rank the pairs, starting with the smallest as 1. Ties receive a rank equal to the average of the ranks they span. Let R_i denote the rank.
5. Calculate the test statistic W

$$W = \sum_{i=1}^{N_r} [\operatorname{sgn}(x_{2,i} - x_{1,i}) \cdot R_i],$$ the sum of the signed ranks.

The signum function of a real number *x* is defined as follows:

$$\operatorname{sgn}(x) := \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$$

Hypothesis Testing: Summary

- Calculate the test statistic
  - Different hypothesis tests are appropriate, in different situations

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Calculate the p-value on the test statistic
- If p-value is "small" then reject the null hypothesis
  - "small" is often $p < 0.05$ by convention (95% confidence)
  - Many data scientists prefer a smaller threshold often 0.01 or 0.001.

Generating a Hypothesis: Type I and Type II Error

If H_0 is X, and we …:	**Null hypothesis(H_0) is true**	**Null hypothesis(H_0) is false**
Fail to accept the Null Hypothesis → we claim something happened	**Type I error** **False positive** **α**	**Correct Outcome** **True positive** **We reject the Null hypothesis**
Fail to reject the null hypothesis → we claim nothing happened.	**Correct outcome** **True negative** **Accept the NULL hypothesis**	**Type II error** **False negative** **β**

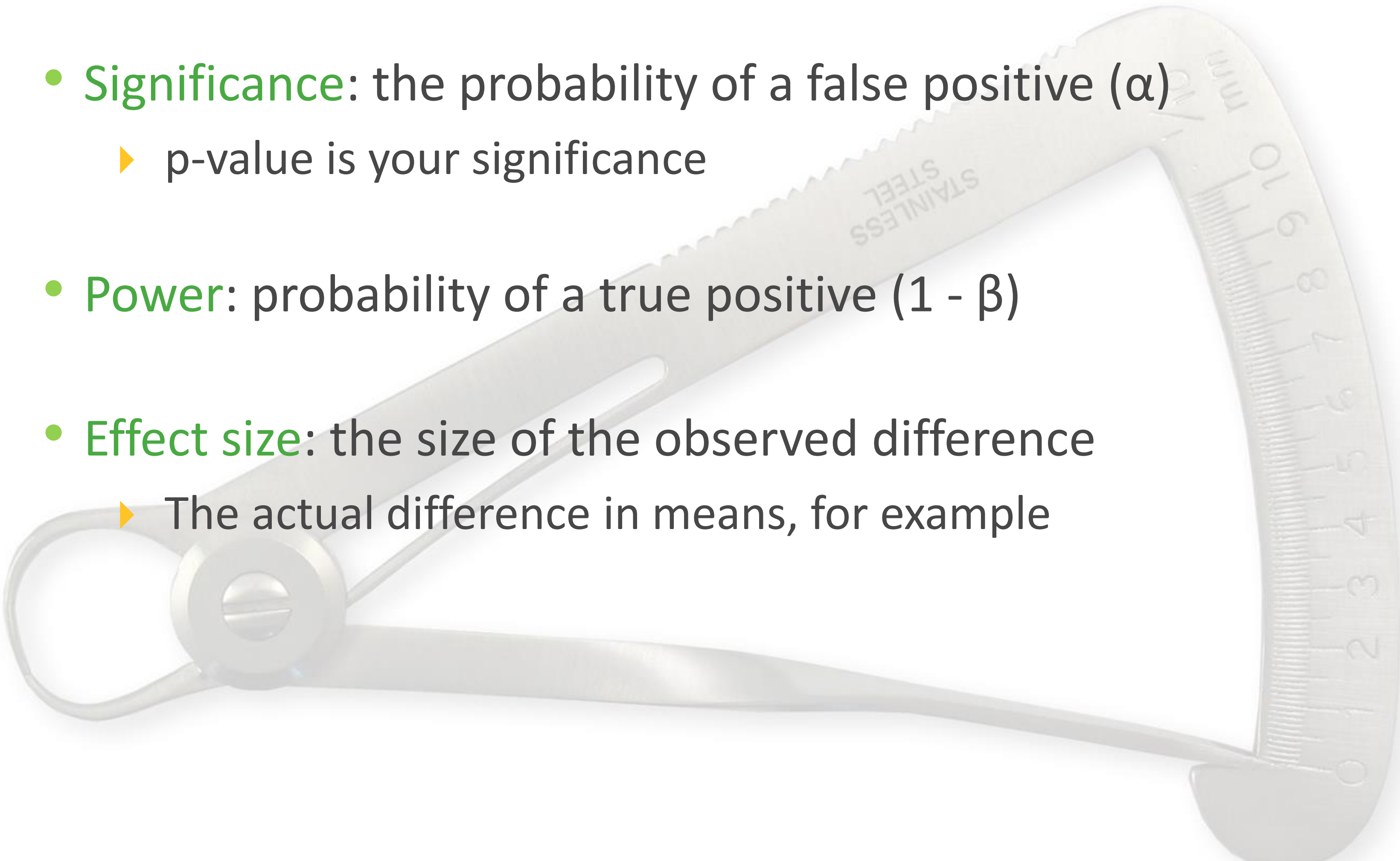
Example: Ham or Spam? H_0: it's Ham H_A: it's Spam

It's Really -> we say it's ↓	**Ham**	**Spam**
Spam	**Type I – false positive**	OK – true positive
Ham		**pe II – false negative**

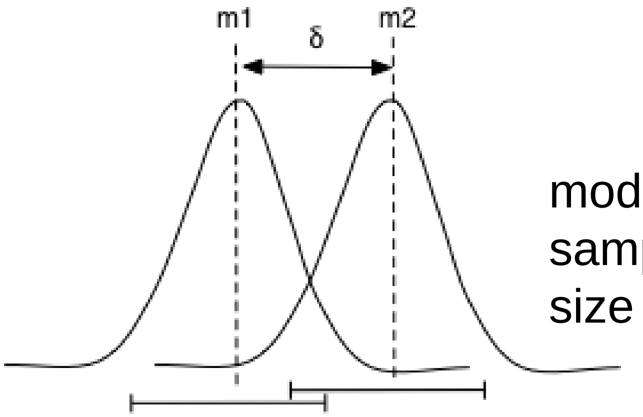
- **Goal: Identify Spam**
- **Which error is worse?**

Significance, Power and Effect Size

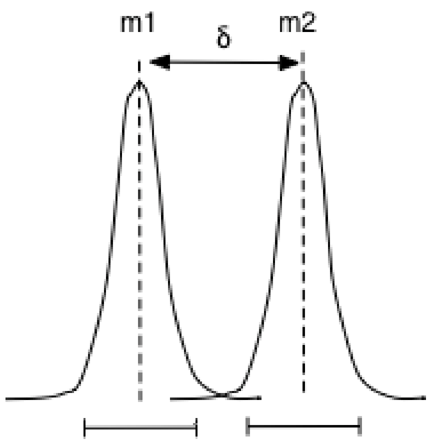
- Significance: the probability of a false positive (α)
  - p-value is your significance
- Power: probability of a true positive ($1 - \beta$)
- Effect size: the size of the observed difference
  - The actual difference in means, for example



Always Keep Effect Size in Mind!



moderate sample size



larger sample size

Both power and significance increase with larger sample sizes.

So you can observe an effect size that is *statistically* significant, but *practically* insignificant!

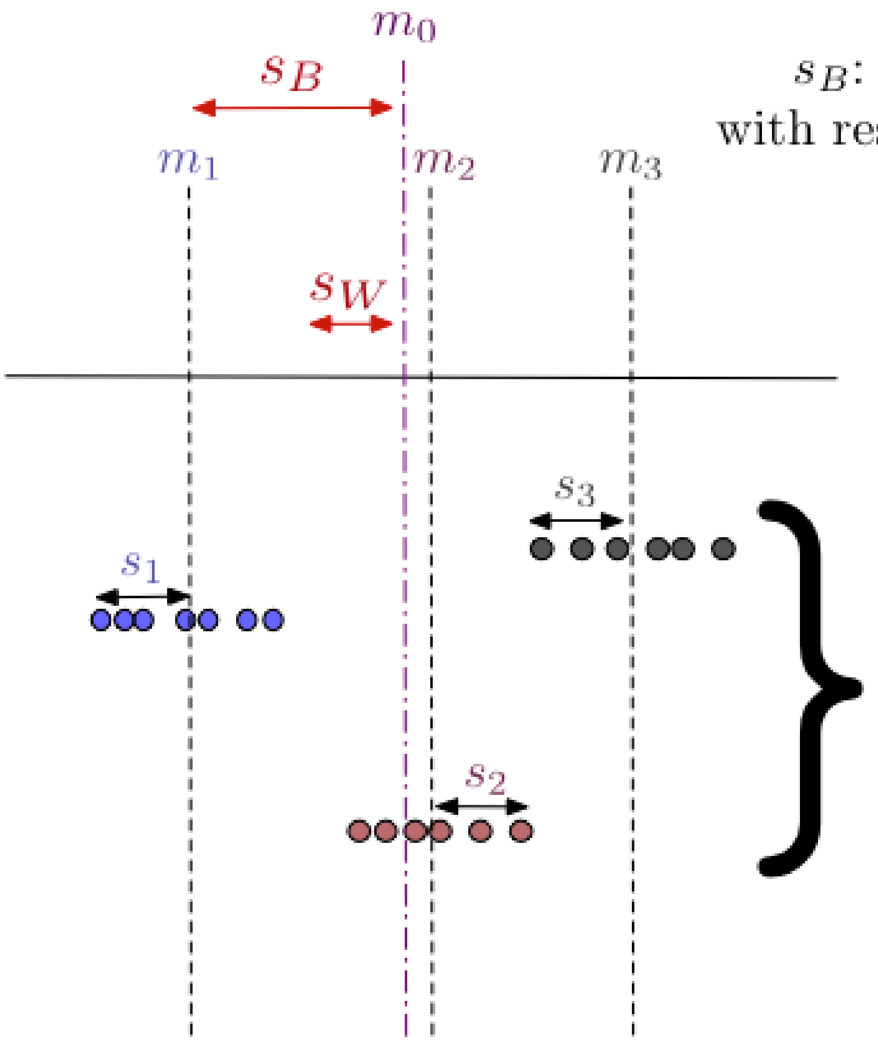
Hypothesis Testing: ANOVA

ANOVA is a generalization of the difference of means

- One-way ANOVA
  - k populations ("treatment groups")
  - n_i samples each – total N subjects
  - Null hypothesis: ALL the population means are equal

Population	n_i: # offers made	m_i: avg purchase size
Offer 1	100	$55
Offer 2	102	$50
No intervention	99	$25

ANOVA: Understanding the F statistic



s_B: how the population means vary with respect to the total mean m_0

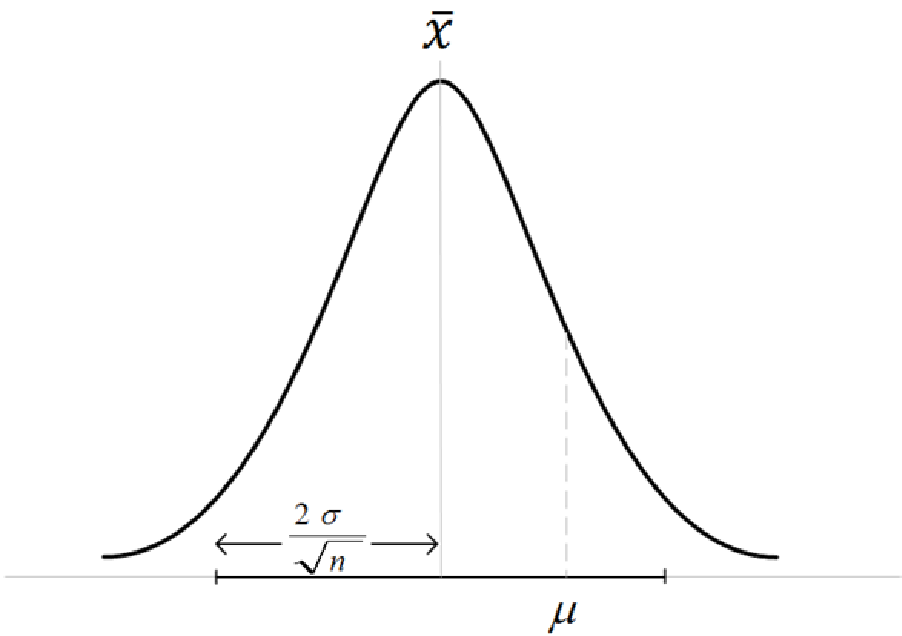
$$\boxed{s_B{}^2 = \frac{1}{k-1} \sum_i n_i \cdot (m_i - m_0)^2}$$

$$\boxed{S_W^2 = \frac{1}{N-k} \sum_i^k \sum_j^{n_i} (x_{ij} - m_i)^2}$$

s_W: the "average" of the s_i

Test statistic: $F = s_B{}^2 / s_W{}^2$

Confidence Intervals



<u>Example:</u>

- Normal data N(μ, σ)
- x is the estimate of μ
  - based on n samples

μ falls in the interval

$$\bar{x} \pm 2\sigma/\sqrt{n}$$

with approx. 95% probability ("95% confidence")

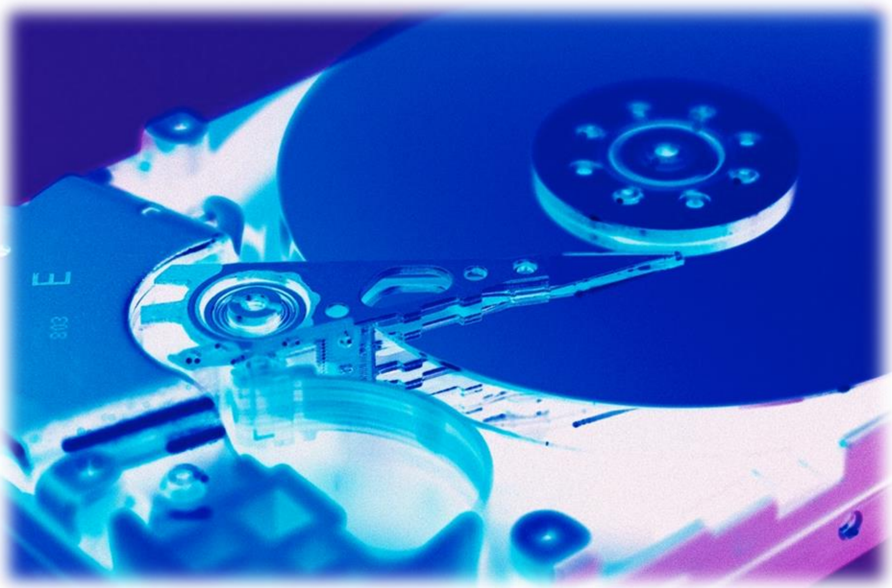
If x is your estimate of some unknown value μ, the P% confidence interval is the interval around x that μ will fall in, with probability P.

Example

The defect rate of a disk drive manufacturing process is within 0.9% - 1.7%, with 98% confidence. We inspect a sample of 1000 drives from one of our plants.

- We observe 13 defects in our sample.
  - Should we inspect the plant for problems?

- What if we observe 25 defects in the sample?



Check Your Knowledge

- Refer back to the ANOVA example on an earlier slide. What do you think? Does the difference between *offer1* and *offer2* make a practical difference? Should we go ahead and implement one of them?

- If yes, and the costs were US $25 for offer1 and US $10 for *offer2*, would you still make the same decision?

- In our manufacturing plant example, assuming you would check the plant for problems in the manufacturing process, how might you justify this decision financially?

Module 3: Review of Basic Data Analytic Methods Using R

part 3: Summary

During this lesson the following topics were covered:

- The role of Statistics in the Analytic Lifecycle
- Developing a model and generating the null and the alternative hypothesis
- Difference between means
- Difference between significance, power and effect size, and how they relate to Type I and Type II errors
- Applying ANOVA and determining whether the results are significant
- Defining confidence intervals and applying them

Lab Exercise 3: Basic Statistics, Visualization and Hypothesis Tests

This lab is designed to investigate and practice using R to perform basic statistics and visualization on data and to perform hypothesis testing.

- After completing the tasks in this lab you should able to:
  - Perform basic data analysis
  - Visualize data with R
  - Create and test a hypothesis

Lab Exercise 3: Basic Statistics, Visualization and Hypothesis Tests– Part1 - Workflow

1. Prepare working environment for the Lab and load data files
2. Obtain summary statistics for Household Income and visualize data
3. Obtain summary statistics for number of rooms and visualize data
4. Remove Outliers
5. Stratify Variable – Household Income and plot the results
6. Plot Histogram and Distributions
7. Compute Correlation between income and number of rooms
8. Create a Boxplot – Distribution of income as a factor of number of rooms
9. Exit R

Lab Exercise 3: Basic Statistics, Visualization and Hypothesis Tests - Part 2 - Workflow

1. Define problem – Analysis of Variance (ANOVA)
2. Generate the Data
3. Examine the Data
4. Plot and determine how purchase size varies within the three groups
5. Use lm() to do the ANOVA
6. Use Tukey's test to check all the differences of means
7. Use the lattice package for density plot
8. Plot the Logarithms of the Data
9. Use ggplot() package
10. Generate the example data to perform a Hypothesis Test with manual calculations
11. Create a function to calculate the pooled variance, which is used in the Student's t statistic
12. Examine the Data
13. Calculate the t statistic for Student's t-test
14. Calculate the degrees of freedom
15. Compute the area under the curve
16. Perform Student's t-test directly and compare the results

Thanks